DOCUMENT RESUME

ED 068 531 TM 001 897

AUTHOR	Hsu, Tse-Chi; Boston, M. Elizabeth
TITLE	Criterion-Referenced Measurement: An Annotated Bibliography.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	LRDC-72-9
PUB DATE	Feb 72
NOTE	25p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Annotated Bibliographies; *Bibliographies; Booklists; *Criterion Referenced Tests; Documentation; *Evaluation Criteria; Instructional Design; *Measurement Techniques

ABSTRACT

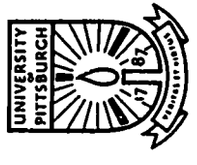
This bibliography is an attempt to assemble and annotate published and unpublished papers, studies, and related articles that have some bearing on criterion-referenced measurement. Conciseness and accuracy are two major criteria in the description of the content of the materials. An attempt was made to avoid critical comment. The 52 items included data from 1913 to 1971, although the majority of them are very recent. The bibliography should be of interest to those working in measurement, individualized instruction, and accountability. (Author/LH)

LEARNING RESEARCH AND DEVELOPMENT CENTER

1972/9

CRITERION-REFERENCED MEASUREMENT: AN ANNOTATED BIBLIOGRAPHY

TSE-CHI HSU AND M. ELIZABETH BOSTON



Criterion-Referenced Measurement: An Annotated Bibliography

Tse-Chi Hsu

and

M. Elizabeth Boston

Learning Research and Development Center
University of Pittsburgh

Abstract

This bibliography is an attempt to assemble and annotate published and unpublished papers, studies, and related articles that have some bearing on criterion-referenced measurement. Conciseness and accuracy are two major criteria in the description of the content of the materials. An attempt was made to avoid critical comment.

It is hoped that readers may gain an initial idea about the content of the papers, and eventually seek out those papers relevant to their interests. This bibliography should be of interest to anyone working in measurement, individualized instruction, and accountability.

Criterion-Referenced Measurement: An Annotated Bibliography

Tse-Chi Hsu

and

M. Elizabeth Boston

Learning Research and Development Center

University of Pittsburgh

February 1972

The preparation of this paper was supported by the Learning Research and Development Center supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in the paper do not necessarily reflect the position or policy of the Office of Education and no official endorsement should be inferred.

Preface

Interest in "criterion-referenced measurement" has been rekindled in recent years, due to the emphasis in education upon such innovations as programmed learning, individualization of instruction, performance contracting, accountability, and computer testing. These innovations often demand an approach in measuring achievement that is different from traditional measurement techniques and one which criterion-referenced measurement might fulfill.

In view of the difficulty of finding related literature for criterion-referenced measurement (CRM) in 1970, we accepted the suggestion of Dr. Anthony J. Nitko to compile a special file for CRM in the Measurement and Evaluation Project of the Learning Research and Development Center. We attempted to assemble published and unpublished papers, studies, and related articles that have direct bearing on CRM. The basis of the collection was papers presented at national professional conventions in the last several years.

After that initial effort, we received many requests from colleagues and friends to recommend appropriate literature for CRM. We decided to abstract the papers in our file and to search for other related articles. The search for articles in educational measurement journals was disappointing. In reviewing the ERIC indices, Research in Education (RIE), 1956-1971 and Current Index to Journals in Education (CIJE), 1969-1971, we did not find too many articles dealing with CRM.

Most of the articles dealing with criterion problems refer to criterion variables alone. This bibliography is concerned with articles that can fit into the interpretation of criterion-referenced measurement which incorporates absolute standards in measuring human behaviors. Sources are indicated for each paper whenever possible. Papers which are not available through any source, except the author, are also included. Readers interested in these papers may have to write to the author for the complete papers. However, there is no guarantee of the availability of such papers, since the authors may not have copies for distribution. Papers with the notation "abstract" are papers which we did not obtain; the annotations were made from the abstracts published by the professional societies.

The two main criteria used in writing the annotations were conciseness and accuracy. An attempt was made to avoid critical comment. It is hoped that readers may gain an initial idea about the content of the papers, and eventually seek out those papers relevant to their interests.

We hope that this first attempt at compiling a comprehensive bibliography will generate interest in CRM, elicit suggestions for improvement of the bibliography, and bring to light other articles for an updated revision of the bibliography in the near future, especially since interest in the research field of CRM is gaining momentum.

We wish to acknowledge our indebtness to Dr. Robert Glaser and to Dr. Anthony J. Nitko for their encouragement and for providing copies of articles from their files. To Mrs. Patricia Graw, we offer our thanks for her fine work and patience in typing this manuscript.

Tse-Chi Hsu
M. Elizabeth Boston

Criterion-Referenced Measurement: An Annotated Bibliography

Angoff, William H. Scales with nonmeaningful origins and units of measurement. Educational and Psychological Measurement, 1962, 22(1), 27-34.

The four reasons usually given for using derived score scales for standardized tests are discussed: 1) convenience of handling test score data, 2) equality of units, 3) equating of forms, and 4) normative meaning. The author disagrees with the normative characteristic of the scales and examines the problems of the normative scale. With the passage of time, the meaning of the normative scale cannot be held as a constant. Therefore, the author suggests that non-normative scales should be used. It is the responsibility of the publisher to provide a variety of current and useful norms, but it is the user who is supposed to incorporate meaning into the scale based on his own experiences.

Bloom, Benjamin S. Learning for mastery. In B. Bloom, J. Hasting, and G. Madaus (Eds.), Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971, 43-57.

This paper does not deal directly with criterion-referenced measurement, but it provides a good rationale for using criterion-referenced measurement. If Carroll's concept, that aptitude is the amount of time required by the learner to attain mastery of a learning task, is useful, what the educator should do is provide enough time for children to learn any skill and use "the formative tests" to measure the mastery or non-mastery of that skill. The formative tests or diagnostic-progress tests referred to in this paper are criterion-referenced measures. The paper also reflects the inadequacy of traditional norm-referenced measurement for measuring mastery of skills.

Blumenfeld, Gerald J., Bostow, Darrel, and Waugh, Robert. Effect of criterion-referenced testing upon the use of remedial exam opportunities. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971. (abstract)

Students in a large educational psychology class were assigned to high criterion and no criterion conditions in regard to passing weekly exams. Counterbalance was used to control sequential effect. The percentage of students not attaining criterion on the initial exam, but taking the remedial exam for that week was higher for the high criterion group in most cases. The percentage of students attaining criterion on either the initial or remedial test was also higher for the high criterion groups.

Cartier, Francis A. Criterion-referenced testing of language skills. Tesol Quarterly, 1968, 2(1), 27-32. ERIC 1968, ED 020 515.

Eight points of contrast between criterion-referenced measurement and norm-referenced measurement are listed; these include the range of scores each type of test is designed to produce (criterion-referenced tests--no range), behaviors tested, how many items can be missed, the importance of which items are missed, what happens to items missed by many subjects, and the security of test items.

The difficulties encountered in applying concepts of criterion-referenced measurement in the development of a program to teach English to foreign military personnel are discussed.

Cleary, T. Anne. Strategies for criterion-referenced test construction using classical procedures. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971. (abstract)

This paper investigates classical procedures and their relationship to both criterion-referenced tests and norm-referenced tests. It endorses the use of classical procedures, such as reliability, discrimination index, etc., in the construction and evaluation of criterion-referenced measures. Strategies for using classical procedures in criterion-referenced measurement are discussed.

Cotton, Timothy S. An empirical test of the binomial error model applied to criterion-referenced tests. Unpublished doctoral dissertation, University of Pittsburgh, 1971.

The major purpose of the study is to provide some empirical evidence for the binomial error model as applied to criterion-referenced tests. Five studies were conducted to investigate: 1) the appropriateness of the binomial model in terms of predicting the distributions of observed scores as a function of operationally defined true scores and test length, 2) the effects of item heterogeneity upon the appropriateness of the binomial model, and 3) the effects of employing item stratification procedures on tests of various lengths.

The study has provided some evidence for the validity of the binomial model, but it is not definitive. The effects of item homogeneity of content on the validity of the model did not appear significant. The effects of using item stratification procedures yielded little differences when dealing with a heterogeneous content domain but seemed beneficial when dealing with a more homogeneous domain.

Coulson, John E. and Cogswell, John F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2(1), 59-64.

The authors are concerned with the impact of programmed learning and individualized instruction on the educational system, especially as it relates to testing and diagnosis in schools. Schools of the future are envisioned as having, primarily, computer-based instruction; two different computer projects are described. The "engineering" approach for research is described and suggested as a more practical research approach at this stage of development than the control-group comparison method, particularly for developing diagnostic tools.

Cox, Richard C. Evaluative aspects of criterion-referenced measures. In W. J. Popham (Ed.), Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 67-75. ERIC 1970, ED 038 679.

A distinction is made between norm-referenced and criterion-referenced measurements in terms of the type of information provided by each, although a single test can yield both norm-referenced and criterion-referenced information. The author discusses an approach to item analysis that appears appropriate for criterion-referenced measures: discrimination between pre- and posttest groups. He also suggests that the coefficient of reproducibility be used as an estimate of reliability across all individuals taking the test.

Content validity in criterion-referenced measurement depends upon the correspondence of items to the specific behavioral objectives. The author calls for increased investigation of alternate approaches to reliability, validity, and item analysis when dealing with criterion-referenced tests and a broader interpretation and application of this kind of measurement.

Cox, Richard C. and Vargas, Julie S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, February, 1966. ERIC 1967, ED 010 517.

Since norm-referenced measures indicate the relative standing of an individual in a group, the traditional technique of item analysis, which selects items that maximize differences among individuals, is an appropriate technique to use with such measures. Criterion-referenced measures, however, focus attention on specific behaviors that an individual masters or does not master, without regard to other individuals. As such, items cannot always be selected by employing the usual item analysis technique.

This study compares the results of the use of two different discrimination indices which were computed for items on tests given in a pre-posttest situation:

1) the common upper/lower 27 percent method.
2) percent passing posttest minus percent passing pretest.

The results indicate that the correlation coefficient between these two indices is significant at the .01 level, but some items acceptable in the pre-posttest technique were eliminated by the usual discrimination index.

Crawford, William R. Assessing performance when the stakes are high. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

The author describes a procedure that was developed at the University of Illinois College of Medicine to measure the achievement levels of medical students. The criterion-related measurement technique seeks to answer the question, "What can this candidate do at a given time under certain circumstances?" The author stresses the value and importance of this type of measurement in the medical profession as opposed to the use of norm-referenced methods. The concept used in scoring examinations, the "Minimum Passing Level," is explained. The use of similar techniques in other professions is suggested.

Davis, Frederick B. Criterion-referenced tests. In the Proceedings of the Thirty-Fifth Annual Conference of Educational Records Bureau. Greenwich, Connecticut: Educational Records Bureau, October, 1970, 40-42.

The author maintains that tests are not properly described as "norm-referenced" or "criterion-referenced"; these two terms should more readily be applied to scores, since either type of score can be established for any test. However, the author would like to see the term "criterion-referenced scores" eliminated, because not only is the term confusing to laymen and educators alike, but it is also easily and frequently misinterpreted. Of the substitute terms suggested and evaluated, "mastery-test score" is judged the most acceptable. "Comparison-score" is the term suggested for "norm-referenced score."

The author presents a capsule history of "criterion-referenced tests," and notes their association with the development of individualized instructional programs. In any situation, however, it is difficult to prevent comparison of a pupil's achievement with other pupils, so that some normative aspects are always present.

This paper concludes with a description of the contribution that test theory can provide for mastery tests in regard to: content validity, interpretation of scores, estimating reliability, evaluating errors, length, format, and scoring.

Ebel, Robert L. Content standard test scores. Educational and Psychological Measurement, 1962, 22(1), 15-25.

In this paper the author defines content standard test scores, describes their close relationship to raw scores, compares them to normative standard scores, and suggests two ways of securing test scores having content-meaning. The stated purpose of the paper is ". . . to emphasize the need for and to demonstrate the possibility of test scores which report what the examinee can do."

Ebel, Robert L. Some limitations of criterion-referenced measurement. In the Proceedings of the Thirty-Fifth Annual Conference of Educational Records Bureau. Greenwich, Connecticut: Educational Records Bureau, October, 1970, 35-37. Also in School Review, 1971, 79(2), 282-288. ERIC 1970, ED 038 670.

In this paper, the author discusses three major limitations of criterion-referenced measurements:

1) They do not tell us all we need to know about achievement, e.g., there is no reference to excellence or deficiency.

2) They are difficult to obtain, e.g., they are based on specific objectives, the formulation of which is impractical.

3) They are necessary for only a small fraction of important educational achievement, e.g., should mastery be the same for all?

He also suggests that the direction educational measurement should take is not toward the development of criterion-referenced measures (which are not new), but toward the refining and improvement of norm-referenced measures.

Ebel, Robert L. Some problems in assessing educational performance. Paper presented at the National Conference on Performance Contracting, Washington, D. C., December, 1971.

In the section, "How should performance be assessed?," the author points out two unrealistic expectations of criterion-referenced tests: 1) most of the students will attain most of the objectives of instruction in a well-taught course, and 2) test items based on a separate important objective of instruction will assess more accurately how much a student has actually learned. Although the purposes and the construction procedures might distinguish norm-referenced tests from criterion-referenced tests, the practical differences between these two types of tests are not substantial. The author concludes that "criterion-referenced testing offers no great promise for substantial improvements in the assessment of educational performance."

Ferguson, Richard L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 81(1), 25-31.

This study shows how a computer can be used to generate, present, score, and interpret criterion-referenced tests. Using Wald's sequential probability ratio test, the computer determined whether the examinee was or was not proficient in the skill being tested. Results show that the computer test was highly valid and reliable. The testing time required to obtain a test profile for the given unit was substantially reduced.

Flanagan, John C. Units, scores, and norms. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951, 695-763.

"Standard of performance" is distinguished from "norm-performance" and regarded as the most fundamental piece of information that an achievement test should provide. Standard of performance is a description of an individual's performance with respect to some defined body of content that can be interpreted without reference to the scores of other individuals or to norm groups.

Flanagan, John C. Discussion. Educational and Psychological Measurement, 1962, 22(1), 35-39.

This paper, as the title suggests, is a discussion of the three papers (Angoff, Ebel, Gardner) presented in a symposium at the annual meeting of the American Educational Research Association in 1960. The author reviews three ways of incorporating meaning into scores, and examines instances where he is in agreement and disagreement with the use of normative standard scores and content standard scores. The question of whether the slow progress in the development of content scores is due to difficulties involved in their development, or lack of need for them, is posed.

Gagné, R. M. Some notes on criterion-referenced measurement. Florida State University, December, 1969. (mimeo)

"What is measured" should be the major issue in the development of criterion-referenced measurement. Therefore, primary attention should be given to the single item. Characteristics of criterion-referenced measurement are: 1) distinctiveness of items in measuring a particular class of performance, 2) freedom from distortion arising from sources other than learning itself, 3) scoring based on the single item rather than a test, 4) inapplicability of the concept of difficulty, since items should be distinctive and free from distortion, 5) establishment of reliability by use of two items only from a single class of behavior, and 6) appropriateness of "content validity" rather than predictive validity.

Gardner, Eric F. Normative standard scores. Educational and Psychological Measurement, 1962, 22(1), 7-14.

In order for a score to have meaning, it must have a frame of reference. Dissatisfaction with the "content frame of reference" led to normative standard scores which provide a more meaningful interpretation of test scores. The author examines five desirable properties of test items, discusses the problems encountered in the sampling of items and examinees, and presents a strong case for the use of different types of norms which yield more useful information than can be gleaned from raw scores alone. These concepts should be kept in mind in dealing with criterion-referenced measurement.

Garvin, Alfred D. The applicability of criterion-referenced measurement by content area and level. In W. J. Popham (Ed.), Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 55-63. ERIC 1968, ED 041 038.

The ultimate purpose of measurement is to provide information for decision making. The question here is not whether to use criterion-referenced measures at all, but when to use them. The author cites examples where criterion-referenced measurement is applicable, and examples where norm-referenced measurement is the appropriate measurement for decision making. Instructional objectives, criteria, and measurement are discussed, and four general principles are advanced as guidelines for applying criterion-referenced measurement to various content areas and various levels of these areas.

Glaser, Robert. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521. Also in W. J. Popham (Ed.), Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 5-14.

The author begins by explaining the difference between aptitude measures and achievement measures. He then discusses the two types of information revealed by scores from an achievement measure, which are dictated by the standard used as a reference: 1) what a student can or cannot do--criterion-referenced, and 2) a student's performance in comparison to group performance--norm-referenced.

The differences encountered in the construction of tests that discriminate among individuals and discriminate between pre- and post-instruction groups are presented. However, in view of the development of instructional technology and individualized instruction, some additional considerations on measurement procedures should be made.

Glaser, Robert and Cox, Richard C. Criterion-Referenced Testing for the Measurement of Educational Outcomes. Pittsburgh, Pennsylvania: Learning Research and Development Center, 1968. (Reprint 41) (Reprinted from Instructional Process and Media Innovation, edited by Robert A. Weisgerber. Chicago: Rand McNally, 1968, 545-550.) ERIC 1970, ED 038 832.

This is a revised version of Glaser's (1963) original paper. Additional attention is given to the implications for achievement test development. Statistical properties of test items are compared in terms of maximizing individual differences in criterion-referenced test items. The authors emphasize that criterion-referenced tests require a different method of construction than do norm-referenced tests.

Glaser, Robert and Klaus, David J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological Principles in System Development. New York: Holt, Rinehart and Winston, 1962, 419-474.

A thorough treatment of the many and varied aspects of proficiency measurements is presented here. These aspects include: 1) characteristics, such as norm-referenced and criterion-referenced; 2) uses, for example, assessing individual differences as opposed to assessing group differences; 3) definition of behaviors to be measured, for instance, identification, quantification, and simulation; 4) sampling and the relative importance of performance components; 5) precision and relevance including contamination, reliability, and validity; 6) eliciting behaviors; and 7) applications. Although the focus of discussion centers upon the measurement of proficiency of the human component in a man-machine system, the suggestions presented should be useful for those dealing with criterion-referenced measurement.

Glaser, Robert and Nitko, Anthony J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1970, 625-670.

A portion of this chapter is devoted to criterion-referenced testing. Consideration is given to the comparison of norm-referenced tests with criterion-referenced tests, the construction of criterion-referenced tests, and the interpretation of test scores.

Hammock, J. Criterion measures: Instruction vs. selection research. American Psychologist, 1960, 15, 435. (abstract)

A distinction is made between tests developed as criteria for selection research and tests developed as criteria for experimental evaluation of instructional programs. An analysis of the desired attributes of instructional research criteria, and a rationale for the development of instructional research criteria are presented.

Harris, Chester W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

The author presents an analysis of Livingston's reliability coefficient for criterion-referenced measures. He rejects Livingston's coefficient since it is the same as a conventional reliability coefficient when that coefficient is based on two populations with means equally distant above and below the criterion score. The author also suggests that the larger Livingston coefficients are secured by extending the range of talent, and that since the standard error of measurement remains the same, a larger coefficient does not imply a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value.

Harris, Margaret L. and Stewart, Deborah. Application of classical strategies to criterion-referenced test construction: An example. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971.

The authors describe a method used in developing criterion-referenced tests to determine mastery of particular skills in an individualized reading program. Test items were administered to a sample of students with a variety of proficiencies in a skill, and classical test theory techniques were then used successfully to refine the tests.

Hills, John R. Experience in small graduate classes and approaches to evaluating criterion-related measures. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

In this paper the author relates the results of a small experiment conducted in his own graduate courses which convinced him that the use of behavioral objectives and course evaluation using criterion-referenced procedures was an "effective way to enhance learning."

The second portion of the paper concerns statistical and item analysis procedures and the interpretation of data for criterion-referenced tests. The author describes the use of the "two difficulty levels" approach (Pre- and Posttest) and advocates the use of stability of score as a measure of reliability. He then considers a way of determining content validity (compare items to the behavioral objective) and "transfer validity" (Are the successful students superior to other students in later courses or activities?).

Hsu, Tse-Chi. Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971.

A good discriminating item for criterion-referenced tests is the one which has a larger proportion of correct responses in the mastery group and a smaller proportion of correct responses in the non-mastery group. Based on this definition, the difference in proportions of correct responses in mastery and non-mastery groups and the phi (ϕ) coefficients are proposed as discrimination indices for criterion-referenced test items. These two indices were compared empirically with the point biserial correlation of items and test scores in three different situations: 1) subjects vary in ability, 2) items vary in difficulty, and 3) score distributions vary in shape.

Jackson, R. Developing criterion-referenced tests. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1970. ERIC, ED 041 052.

The current definitions of criterion-referenced testing are inadequate. Criterion-referenced tests should be produced by objectively defined processes. Principles for developing criterion-referenced tests are advanced, and difficulties such as objectivity, reproducibility, and generalizability are explored. The author expresses doubt as to the development of criterion-referenced tests for complex behaviors. He also examines the value of item analysis, scalability, reliability, and validity as processes for empirically evaluating criterion-referenced tests.

Klein, Stephen. Evaluating tests in terms of the information they provide. Evaluation Comment, Center for the Study of Evaluation, UCLA, 1970, 2(2), 1-6. ERIC 1971, ED 045 699.

Criterion-referenced measures and norm-referenced measures are compared in terms of the purpose of the tests and the philosophy underlying the manner in which the tests are constructed. The author points out advantages and limitations of using norm data and criterion data in evaluating student performance and in program evaluation. In view of the difficulty in using traditional test construction procedures to accomplish both purposes, the author suggests a technique of combining the better components of the norm- and criterion-referenced approaches. The essential component of this technique is to include the concept of item difficulty and normative score reporting in the development and interpretations of criterion-referenced measures. Examples are provided to illustrate this new procedure.

Kriewall, T. E. and Hirsch, E. The development and interpretation of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, California, February, 1969. ERIC 1971, ED 042 815.

Following the discussion on properties of criterion-referenced tests, the paper introduces a model for developing and interpreting criterion-referenced tests. The model is based on a binomial distribution by regarding the examinee's performance on a test as a series of independent Bernoulli trials. According to this model, the error of measurement is a function only of test length and the examinee's proficiency.

The paper also discusses techniques of quality control using criterion-referenced tests and procedures for minimizing test length. The Neyman-Pearson theory of hypothesis testing and Wald's sequential probability ratio test are possible approaches in reducing the testing time needed to detect mastery levels. The relationships between mastery criteria and various sampling plans, such as single sampling and simple curtailed testing, are also discussed.

Lindvall, C. M. and Nitko, A. J. Criterion-referenced testing and the individualization of instruction. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, California, 1969. ERIC 1970, ED 036 167.

A basic element in achievement testing and in the evaluation of achievement is the determination of whether an individual can or cannot perform some specific skill. The type of information realized from such achievement testing is criterion-referenced test information, which requires a clear description of the performance being assessed. In some situations, criterion-referenced scores can be derived from criterion-referenced information, but the emphasis here is on securing criterion-referenced information. The authors of this paper distinguish between the terms criterion-referenced information, criterion-referenced meaning, and criterion-referenced scores. They also examine the differences between criterion-referenced tests and norm-referenced tests. The use of criterion-referenced testing in the Individually Prescribed Instruction (IPI) Math program is presented as an example of the application of the rationale developed in this paper.

Livingston, Samuel A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

The author presents a theory of reliability for criterion-referenced tests which he has developed. The theory is based on the assumptions of classical test theory. The criterion score is substituted for the mean score of a norm group, and the variance, covariance, and correlation are redefined. The resulting criterion-referenced reliability will be at least as large as the norm-referenced reliability, and when the mean score is equal to the criterion score, the two reliability coefficients will be the same. Implications of this criterion-referenced reliability are also discussed.

Majer, Kenneth and Shoemaker, David M. A three part test for criterion-referenced assessment. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

This paper discusses the construction of trident tests for a third grade spelling class and the results of administration of the tests over a six week period.

The tests consist of three parts:

1) posttest on current material which indicates the effectiveness of instruction.
2) posttest on previously learned material which suggests any need for remedial work.
3) pretest on future material which enables a teacher to prescribe an appropriate path an individual might follow in future work.

The authors feel that more meaningful information can be obtained from tests constructed in this manner than can be had from the usual posttest which tests only current material.

Mattson, Dale E. Criterion related measures in education - an appealing delusion. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

Based on decision theory models, discussion centers on the relationship among the minimum mastery level, the needs of society, and the cost of training. The minimal pass level is "set in such a way as to best meet the needs of society at a price society can and will pay." Moreover, the author asserts that all standards of evaluation for the established professions of the United States exceed absolute minimal standards by a wide margin. Therefore, the performance of each student has to be evaluated against the performance of a norm group, a group of potential competitors who wish to provide a service to society at a cost which society is willing to pay.

Millman, Jason. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52(4), 226-230.

This article deals with reporting school progress to students and parents in an individualized instructional setting where criterion-referenced measurement is the appropriate method of assessment. There is a short discussion on two ways of individualizing: pacing and branching, and a comparison of criterion-referenced measures with norm-referenced measures. The author cites some of the advantages and disadvantages for a school when a program requiring the use of objectives, individualization, and hence, criterion-referenced testing is introduced. A sample report card for use with such a program is presented here.

Nitko, Anthony J. Criterion-referenced testing in the context of instruction. In the Proceedings of the Thirty-Fifth Annual Conference of Educational Records Bureau. Greenwich, Connecticut: Educational Records Bureau, October, 1970, 37-40. (Also LRDC Publication 1971/1) ERIC 1971, ED 047 010.

A brief background of criterion-referenced testing is presented. The paper also considers the relationship between norm-referenced information and criterion-referenced information, and the need for vigorous, empirically based construct validation studies of criterion-referenced tests. However, whether criterion-referenced testing and/or norm-referenced testing is needed to make instructional decisions depends upon the instructional context within which one operates.

Nitko, Anthony J. A model for criterion-referenced tests based on use. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971. ERIC 1971, ED 049 318.

The purpose for which a test is used determines how the test should be designed. Traditional procedures may be applied in the design of criterion-referenced tests in some instances, but must be avoided in others. The differences between cut-off scores, criterion scores, and mastery scores are discussed.

Popham, W. James. Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

This book is a collection of papers presented at the AERA/NCME Symposium (1970) on Criterion-Referenced Measurement: Emerging Issues, including Garvin's "The Applicability of Criterion-Referenced Measurement by Content Area and Level," Cox's "Evaluation Aspects of Criterion-Referenced Measures," and Popham's "Indices of Adequacy for Criterion-Referenced Test Items." In addition to these three papers, the book also includes Glaser's "Instructional Technology and the Measurement of Learning Outcomes: Some Questions" and "A Criterion-Referenced Test," plus Popham and Husek's "Implications of Criterion-Referenced Measurement." A historical introduction is presented by William Trow, and Selected References were prepared by Leonard L. Streeter.

Popham, W. James. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 79-98.

This study attempts to identify useful indicators for criterion-referenced items. The Cox-Vargas analysis was replicated unsuccessfully. A four-fold table representing possible pretest-posttest performance on test items was also investigated. A possible approach suggested by the author is the use of chi-square to contrast the pre- and post-instruction relation of each item with hypothetical frequencies based on the median value of each subtest.

Popham, W. James and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9. Also in W. J. Popham (Ed.), Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971, 17-37.

The authors explore the similarities and differences between criterion-referenced and norm-referenced approaches to measurement in terms of purpose and use. They also distinguish between the two types of approaches by examining several important measurement constructs, especially as they relate to criterion-referenced measures: variability, item construction, reliability, validity, item analysis, reporting, and interpretation.

Rahmlow, Harold F., Matthews, Josephine J., and Jung, Steven M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the joint session of the AERA/NCME annual meeting, Minneapolis, Minnesota, March, 1970.

This paper describes an item analysis technique used in the PLAN program. A criterion-referenced test consisting of easy, middle difficulty, and hard items was administered to two groups of sixth grade students; one group had instruction (post-instruction) and one did not (non-instruction). Item analyses were performed. From the results it appears that Kuder-Richardson 20 is not a useful item statistic for use with criterion-referenced tests, since it relies heavily on item variance. However, the combined use of the difficulty index and the non-instruction to post-instruction gain scores seems to be a satisfactory and useful technique for selecting items for criterion-referenced tests.

Richards, James M. Assessing student performance in college. ERIC 1970, ED 040 307.

Three areas of current research in assessing student performance are reviewed: the development of examinations for which academic credit is awarded, criterion-referenced testing, and the assessment of extra-curricular activities. The author presents a thorough "Overview" and "Technical Review" for each of the three areas.

The idea of criterion-referenced measurement has been around for a long time. Although currently CRM is still a theoretical possibility and not a usable procedure, it is hoped that the idea will not be abandoned without further exploration. Item writing, pooling and analysis, reliability, and the significance of criterion-referenced testing for measurement theory are discussed, along with other problems connected with the construction of such tests.

Roudabush, Glenn and Green, Donald Ross. Some reliability problems in a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, 1971.

The authors describe a "Prescriptive Mathematics Inventory" which offers diagnostic and prescriptive information to the teacher by the use of a limited number of test items rather than an extreme amount. The problems of establishing reliability of criterion-referenced tests in terms of stability of individual performance on single items and of the stability of patterns of right and wrong responses are discussed.

Shoemaker, D. M. Criterion-referenced measurement revisited. Educational Technology, March, 1971.

This paper emphasizes the close relationship between criterion-referenced measurement and an instructional program. A criterion-referenced test is usually one of many tests in a sequence for an instructional system. A test item for a criterion-referenced test does not represent only the behavior indicated by the item, but each item is a subset of all possible items for an objective, which represents a hypothetical content population. In other words, the items for a criterion-referenced test constitute a sample from the content population of testable items.

Three types of items should be included in the criterion-referenced test for each objective: 1) items that can be answered correctly by all students with minimum satisfactory performance, 2) items that can be answered correctly only by students who have surpassed the minimum achievement, and 3) items that can be answered correctly only by students with a high level of achievement. Criterion-referenced tests derive their utility from the meaningfulness and usefulness the information has for the teacher, who formulates the instructional sequence.

Simon, George B. Comments on "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement, 1969, 6(4), 259-260.

The author comments on the Popham and Husek paper reviewed earlier in this bibliography. He contends that the terms norm-reference and criterion-reference refer to test scores, not to the test content. Item sampling, reliability, item analysis, negatively discriminating items, and Guttman reproducibility as presented by Popham and Husek receive consideration in this paper.

Thorndike, E. L. Original tendencies as ends: Emulation in the case of school "marks." A section in author's Educational Psychology. New York: Teachers College, Columbia University, 1913, 1, 286-289.

This is one of the earliest articles related to criterion-referenced measures. Discussion centers on school marks and their meaning. Marks should not express degrees of relative attainment, but should represent objectively defined amounts of knowledge, power, or skill. Thus, a student could monitor his own progress and compete with his own past record, rather than competing with other students.

Thorndike, E. L. The estimation of test validity: Criteria or proficiency. Chapter 5 in author's Personnel Selection Test and Measurement Techniques. New York: John Wiley and Sons, Inc., 1949, 119-159.

An extensive discussion about criterion measures is presented here, with emphasis upon the crucial role of the "criterion" in programs of research for personnel selection and classification. Qualities of criterion measures: reliability, validity, objectivity, and practicality are examined. Several types of criterion measures are discussed in terms of their advantages and disadvantages, and the difficulties inherent in establishing valid, reliable measures are emphasized. Examples are drawn from Air Force training programs.

Unks, Nancy J. An investigation of validity and reliability concepts for criterion-referenced measurement. Unpublished master's thesis, University of Pittsburgh, 1969.

In which ways might traditional measurement theories and techniques be applied to criterion-referenced measurements? This question, with emphasis on reliability concepts and their application, is the major question explored in this logical investigation.

After examining a number of possibilities in each area, the author suggests for use with criterion-referenced measures:

1) the concurrent use of criterion-referencing and sequential scaling for item analysis.

2) construct validity, selective correlation procedures, and an accuracy-of-placement validity for validation.

3) two new reliability concepts: reliability as an indicator of the consistency of the criterion, and reliability as an indicator of the consistency of test validity.

Ward, J. On the concept of criterion-referenced measurement. British Journal of Educational Psychology, 1970, 40, 314-323.

The paper compares norm-referenced measurement and criterion-referenced measurement in terms of inter-subject variability. The author suggests three areas of application for criterion-referenced tests: 1) curriculum evaluation, 2) the sampling of developmental levels in cognitive growth, and 3) the construction of tests for diagnostic/remediation programs in special education. The issues of criterion selection, item selection, and reliability are discussed in light of individual item-sampling procedures.

The author concludes that it is the test user rather than the constructor who postulates criteria in terms of scores obtained. Though there will always be a need for norm-referenced measurement, the author feels the most important future psychological work is in "the identification of meaningful learning criteria, and the accurate assessment of the individual subject's performance with respect to these."

Warrington, Williard G. Criterion related measures: Some general considerations. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minnesota, March, 1970.

The paper begins by citing recent comments concerning testing theory and different views on measuring educational outcomes. The author feels that the emphasis on the differences between criterion-referenced measures and norm-referenced measures will not lead to a new approach in educational measurement, but rather to a different way of assembling test items and using test data. In this frame, the similarities of the two types of measurement devices are stressed, especially the difficulties encountered by both in construction of good test items, specification of educational objectives, and the establishment and agreement on precise criteria for attainment.